Historic, archived document

Do not assume content reflects current scientific knowledge, policies, or practices.

1.9
Ec752Si

Size of Sample Study
Pig Survey

2 Copies - Copy 1

S I Z E O F S A M P L E S T U D Y

With particular reference to the Pig Survey of the

Bureau of Agricultural
Economics

by

Bradford B. Smith
&
Mordecai Ezekiel

Completed October 30, 1924

Contents.

	page
Introduction	1
Experimental data used	1
Theory	2
Application to data	4
Test of Application	5
Representativeness of Sample--Pig Survey	7
Propagation of Error in Obtaining U.S. Average.....	8
Determining the requisite number of records for a given degree of accuracy.....	11
Note on relationship of correlation between x & y in the sample and the correlation between averages of x and y in successive samples.....	13

The authors wish to express their appreciation for the assistance rendered by Mr. H. R. Tolley in the development and application of the theoretical considerations involved.

SIZE OF SAMPLE STUDY--PIG SURVEY.

1) Introduction.

A considerable portion of economic forecasting is based on the "this year to last year" type of ratio. Using the semi-annual "Pig Survey" of the Division of Crop and Livestock Estimates, Bureau of Agricultural Economics, as an example, the reported number of sows farrowed this year and the reported number of sows farrowed last year are tabulated and the ratio between them taken as the ratio of the hog crop to come this year as compared with the known hog crop last year. The occasion for this study was to test the statistical accuracy of this "Sows farrowed this year to sows farrowed last year" ratio. To accomplish this required the application of some known theory and the development of some new. This paper accordingly aims (1) to set forth the theoretical considerations in measuring the accuracy of such a year-year ratio, (2) test it by application to a portion of the records used in making a pig survey, (3) draw certain conclusion specifically related to the pig survey.

2) Experimental data used.

The 8000(cir.) records for the State of Iowa used in the June, 1924 Pig Survey were used as the basis for securing the necessary statistical constants for use in the formulae applicable to the pig surveys and also as experimental material in testing out the application of the theory developed. These records ~~are~~ were secured through the cooperation of the Bureau of Agricultural

Economics and the Post Office Department, the schedules being filled out by rural postmen and returned to the Bureau of

3) Theory

According to definition the standard error of an average, a ratio--indeed of any measure--is the standard deviation of the frequency curve which would be formed by an infinite repetition thereof. That is, if we had a large number of pig surveys taken under identical sampling conditions and the corresponding sows farrowed ratio (X_1/X_2) these ratios would fall on a frequency curve of standard deviation, s; and since the standard deviation of the frequency curve is the standard error of the individual observation, s would then be the standard error measuring the precision of the single ratio, X_1/X_2, or (Sows farrowed this Spring)/(Sows farrowed last Spring). Since we do not have the requisite number of pig surveys to give us the constants, we must approximate them in some other manner.

From Yule 1/, Bowley 2/, and Merriman 3/ the following

1/ G. Udney Yule: Introduction to the Theory of Statistics, p. 215
2/ Arthur L. Bowley: Elements of Statistics, p. 319
3/ Mansfield Merriman: Methods of Least Squares, p. 79

formula for the standard deviation, s, of a series of ratios of form X_1/X_2 may be secured:

$$s^2 = \frac{M_1}{M_2}\left(v_1^2 - 2r_{12}v_1v_2 + v_2^2\right) \qquad \ldots.(1)$$

in which (Yule's terminology), $v_1=\sigma_1/M_1$, $v_2=\sigma_2/M_2$, and M are means of X.

This formula, it may be observed, calls for constants derived from a series of ratios--surveys--while we have but one. The necessary constants being means, standard deviations, and correlation, they may be approximated as follows:

For the ratio of the means of a series of sample pig surveys we may take as our closest approximation the ratio of the single survey which we have. (This ratio in practice never varied beyond .8 - 1.2) The standard deviation of a series of means, is, of course, approximated by the standard error of the single mean and may therefore be secured by dividing the standard deviation in the single sample by the square root of the number of items composing the sample. We may thus approximate v_1 by writing $v_1 = \sigma_1 / (m_1\sqrt{n})$ in which σ_1 is now the standard deviation of the number of sows farrowed this spring, m_1 the mean of the same series and n the number of records in our sample. 1 Similarly $v_2 = \sigma_2 / (m_2\sqrt{n})$

r_{12}, the correlation of a series of averages, m_1 s, with corresponding averages, m_2s, would ordinarily be taken as equal to the correlation between x_1 and x_2 within our sample. But as shown in section nine of this paper this is true only when the relationship between the variates is strictly linear. When the relationship is curvilinear a better measure of the correlation between averages of successive samples is the correlation index 4/

4/ See: Frederick C. Mills: The Measurement of Correlation and the Problem of Estimation. Journal Am. Stat. Assoc., Vol XIX, No. 147 (Sep 1924) See also: Mordecai Ezekiel: A Method of Handling Curvilinear Correlation for any Number of Variables. Journal Am. Stat. Assoc., VolXIX, No. 148 (Dec 1924).

As a matter of safety, therefore, it is better to use the index. A plotting of the original data revealed a very distinct curvilinear relationship in our experimental data. With last spring as the dependent the correlation ratio was .86 and the correlation index computed from a free hand curve after the method of Ezekiel 4/ was approximately .8 , as compared with a correlation coefficient of .56 .

With these considerations in mind the standard error, e, of the ratio of pigs farrowed this spring to pigs farrowed last spring, m_1/m_2, may now be written

$$e^2 = \left(\frac{m_1}{m_2}\right)^2 \left(\frac{\sigma_1^2}{m_1^2 n} - 2\rho_{2.1}\frac{\sigma_1\sigma_2}{m_1 m_2 n} + \frac{\sigma_2^2}{m_2^2 n}\right)$$

or, bring n outside the parentheses, and take root,

$$e = \left(m_1/m_2\right)\frac{1}{\sqrt{n}}\left(\frac{\sigma_1^2}{m_1^2} - 2\rho_{2.1}\frac{\sigma_1\sigma_2}{m_1 m_2} + \frac{\sigma_2^2}{m_2^2}\right)^{1/2} \quad \ldots(2)$$

4) Application to data.

From the sample data (State of Iowa) considered typical of the Corn Belt, the following values for the necessary constants were secured.

Means.	Standard Deviations.	Correlation
		Coefficient, r_{12} = .56
m_1 = 11.81	σ_1 = 10.45	Ratio , $\eta_{2/}$ =.86
m_2 = 14.38	σ_2 = 13.47	Index , $\rho_{2.1}$, =.8

Substituting these values in the formula (2), we have

$$e = \left(m_1/m_2\right)\frac{.579}{\sqrt{n}} \quad \ldots(3)$$

Another method of stating this relation is "The percentage that the probable error is of the ratio is

$$e' = .6745\,\frac{57.9}{\sqrt{n}} \quad \text{or}$$

$$e' = \frac{39.1}{\sqrt{n}} \quad \ldots(4)$$

Formula (4) is graphed on page____.

5) Test of Application

A test of the theoretical results was made as follows:

(a) The 8000 (cir.) records were shuffled and 30 samples of approximately 290 items were dealt out. The following data for each sample was then recorded:

Number in sample, n
Average number of sows farrowed this spring, m_1
Average number of sows farrowed last spring, m_2
The ratio, m_1/m_2, was computed.

(b) The cards were re-shuffled and similar data recorded for another 30 samples, making 60, N, samples in all.

(c) The following means and standard deviations of this set of N samples were then secured.

$$M_{m_1} = 11.81 \quad M_{m_2} = 14.38 \quad M_{m_1/m_2} = .822$$

$$\sigma_{m_1} = .96 \quad \sigma_{m_2} = 1.24 \quad \sigma_{m_1/m_2} = .050$$

The correlation between the series of averages, m_1 & m_2, was .76 which, it may be noted, approximates $\rho_{2.1} = .8$, the value used in the formulae; the reason for using the index of correlation is given in section nine.

(d) For five of the samples (selected at random) the following values were secured:

	Sample #	σ_1	σ_2	m_1	m_2	r_{12}
$n = 289$	1	11.75	13.96	12.7	16.0	.67
	2	13.99	14.41	12.9	15.2	.50
	3	9.89	17.18	12.7	16.7	.57
	4	8.40	16.00	12.4	14.8	.50
	5	8.98	11.28	11.0	12.6	.52
	Average -	10.60	14.56	12.3	15.0	.55

It may be noted that the correlations in the samples come quite close to the correlation for the state, .51, which is a practical demonstration of the statement made in section nine to the effect

that correlation in the sample indicates correlation in the universe.

(e) Should our hypothesis be correct as stated on p.____ and demonstrated in section nine, and should the data follow a normal distribution, we should be able to arrive at the standard deviation of the 60 average ratios, of type m_1/m_2, by use of the formula (2) using constants derived from the five samples analyzed in detail. That is, (using the value of $\rho_{2.1}$ for the state)

$$e = \frac{12.3}{15.0} \cdot \frac{1}{\sqrt{289}} \left(\frac{10.60^2}{12.32^2} - 2.8 \cdot \frac{10.60 \cdot 14.56}{12.32 \cdot 15.00} + \frac{14.56^2}{15.00^2} \right)^{1/2}$$

$$= .028$$

The error so derived, .028, falls considerably short of the computed error, .050. Before attributing this discrepancy to errors in the theory let us examine the data with a view to discovering its normalcy. If normal distribution exists, or for that matter a skew distribution so long as it is not too radical, the standard deviation of the series composed of means should be approximated by the standard error of the single average. As a matter of fact the figures are thus:

$$\frac{\sigma_1}{\sqrt{n}} = .62 \; ; \; \sigma_{m_1} = .96 \qquad \frac{\sigma_2}{\sqrt{n}} = .86 \; ; \; \sigma_{m_2} = 1.24$$

This disparity naturally diminishes the size of the numerator in the formula and hence reduces the computed error to .028. In this case with means of about the same size as the standard deviations, the distribution must necessarily be far from the normal upon which the error formulae are based: for the lowest limit possible in a record ~~a record~~ is zero, or one standard deviation less than the mean, while there is practically no limit to the highest possible value in a record. A frequency polygon revealed the skewness of

the distribution: some of the larger items were fifteen standard deviations greater than the mean.

Since the estimated error was about one-half the actual, it would be a measure of safety to double the estimated error as an approximation of the probable accuracy--or, conversely, when estimating the number of items to secure any given error in a pig survey ration, one-half this given error should be used in the computations.

6) Representativeness of Sample--Pig Survey.

Two criterea of representativeness of sample were considered:

(a) Whether or not the questionnaire will be returned is largely dependent upon the interest of the informant. Hypothetically the large scale hog producer--owing to his progressiveness and financial ability might well be in the lead in adjustment movements. At the same time his progressiveness would make him more willing to return the questionnaire than smaller producers. Thus the results of the questionnaire might reveal a more progressive situation than actually existed. If this type of bias exists it may be detected by noting if there is any constant relation between size of farm and the ratio change. About three hundred records were accordingly chosen at random and the "Sows farrowed this spring to sows farrowed last spring" ratio computed. The ratio was then correlated with the size of farm. The correlation coefficient was less than .05 showing that there is practically no error attributable to automatic selective sampling in this survey. Another way of stating the conclusion is: It makes no difference whether or not

records are taken from large farms or from small. This conclusion should not be generalized to other schedule inquiries, however, for there are certain features in the collection of the pig survey, peculiar to it alone, and to which the validity of this conclusion might be traced.

(b) The second factor considered with reference to representativeness of sample was the influence of geographical location on the ratio. Put as a question: Does the geographical location have influence upon the Sows farrowed this spring to sows farrowed last spring ratio?" This was answered in the affirmative as would be expected, by computing the ratio for different geographical areas. The ratio was computed for each of the nine crop estimate districts of Iowa and ranged from .79 to 1.13 . The conclusion is that only when the ratios are carefully weighted with reference to the number of sows in the district is the average valid.

7) Propagation of Error in Obtaining U. S. Average.

The United States final figure is obtained by securing a weighted average of the states' ratios for "Sows farrowed this spring to sows farrowed last spring. The unit is the state. The weights are intended to approximate the number of sows farrowed in each state. There is, therefore, a possibility of error in the failure of the selected weights to conform to both the years from which the ratio is computed. This type of error, however, will not be considered in this paper.

The error of a sum in terms of the errors of its parts, with weights, w, is 5/

$$E_s^2 = w_1^2 e_1^2 + w_2^2 e_2^2 + + \qquad \ldots (5a)$$

5/ Bowley, p. 316. Merriman, Chap VII, Formula 102.

Since to obtain the weighted mean, we divide the weighted sum by the sum of the weights, the error of the mean would be ~~the value of~~ obtained from formula (5) ~~divided by~~ times $\frac{1}{(\Sigma w)^2}$. Hence,

$$E_m^2 = \frac{E_s^2}{(\Sigma w)^2} = \frac{w_1^2 e_1^2 + w_2^2 e_2^2 + +}{(\Sigma w)^2} \qquad \ldots (5)$$

To obtain the error for the U. S. Ratio, then, compute the errors for each component of the average by formula (2), page __ , From these errors compute the error of the final average, according to the weighting used, by the above formula, (5).

This, however, necessitates a considerable amount of computation in securing standard deviations and correlation indices, particularly when something like 100 000 records are involved. Since the computation of errors gives us at best, but approximations,

predicated upon an infinite number of events, a condition which actually never obtains, any method whereby we can secure such an approximation without such a large amount of computation would be useful. This may be done by assuming that the probable percentage error for all states may be obtained from formula (4). This is on the hypothesis that rho, and the dispersions, $\frac{\sigma}{m}$, will tend to remain constant. This is not, of course the exact case, but it is rational to suppose that as the mean increases--i.e. as we deal with larger farms--so also will the standard deviation increase; that is, the larger farm is but a magnification of the smaller in its chief characteristics. This being so, the dispersion, $\frac{\sigma}{m}$, will tend to remain constant. The value of rho is a reflection of the rapidity with which conditions may change within the industry; that is a large number of sows on the farm this year, means that there are probably a large number last year. It is probable, therefore, that the value of rho will not change materially since it is tied back rather definitely to the size of farm. And this latter does not change very rapidly.

The formula on this basis, then, with s representing probable percentage error becomes

$$S^2 = \frac{w_1^2 \frac{39.1^2}{n_1} + w_2^2 \frac{39.1^2}{n_2} + \quad}{(\Sigma w)^2} = \frac{39.1^2}{(\Sigma w)^2}\left(\frac{w_1^2}{n_1} + \frac{w_2^2}{n_2} + +\right) \ldots (6)$$

If now we further assume that $w = n$, i.e. that the distribution of the tabulated records is made proportional to the weighting scheme, thus eliminating the necessity of weighting by simple averaging of all records, we may write

$$s^2 = \frac{39.1^2}{(\Sigma n)^2}(n_1 + n_2 + \quad) \qquad \ldots(7)$$

or, $$s^2 = \frac{39.1^2}{(\Sigma n)^2}(\Sigma n) \qquad \ldots(8)$$

or $$s = \frac{39.1}{\sqrt{N}} \qquad \ldots(9)$$

where $N = \Sigma n$, which is identical in form with formula (4).

8) Determining the requisite number of records for a given degree of accuracy.

The final ratio for the U. S. is rarely interpreted to a finer degree than five per cent. This is perfectly proper since factors operating subsequently to the taking of the survey, such as hog cholera, for example, can easily cause a digression of this degree from what the ratio would indicate.

Let us assume, however, that a statistical accuracy of one per. cent must be secured. Dividing this by two--in conformity with the suggestion made on page ___ --gives 0.5%. Now the chance of an average being more inaccurate than four times its probable error, owing to the probabilities of sampling, is but one in one hundred. If the probable error of our ratio was but one quarter of 0.5% or 0.125% it is a practical certainty then that the final ratio would be within one per cent of the true value. Taking 0.125 as the desired percentage probable error we must secure then, in our final ratio, we read from the curve the number of records necessary to secure this and find it to be about 100,000.

In a similar manner if we start with the assumption that we must secure an accuracy assuredly within but 5%, the number of

records required is only about 4,000--a rather startling comparison .

50,000 records assures us an accuracy of about one and one-half percent., (1.4%)

A proposition deserving consideration is: Is the increase in labor for tabulating 100 000 instead of 50 000 justified by the increase of assured statistical accuracy of from 1.4% to 1.0%?

9) Note on relationship of correlation between x & y in the sample And the Correlation Between Averages of x and y in successive samples.

The relation between correlation of x_i and y. in the sample and correlation of a series of successive samples' averages is thus:

We may first generalize from the correlation within the sample, r_{xy}, to the correlation in the universe, R_{xy},

$$r_{xy} = R_{xy} \qquad \ldots(10)$$

within the limits of fluctuation of sampling.

And in like manner, as our best available measures, the standard deviations of the sample are taken as equal to the standard deviations of the universe, σ_x , σ_y .

The regression equation gives the most probable values of x, x', associated with y, assuming linear relationship.

$$x' = R_{xy}\frac{\sigma_x}{\sigma_y}\,y \qquad \ldots(11)$$

or letting $R_{xy}\frac{\sigma_x}{\sigma_y} = b$, $x' = by$ $\qquad \ldots(12)$

From (11)

$$\sigma_{x'} = R_{xy}\frac{\sigma_x}{\sigma_y}\,\sigma_y = R_{xy}\sigma_x \qquad \ldots(13)$$

Thus

$$R_{xy} = \frac{\sigma_{x'}}{\sigma_x} \qquad \ldots(14)$$

From (12)

$$\sigma_{x'} = b.\sigma_y \qquad \ldots(15)$$

Substituting (15) in (14)

$$R_{xy} = \frac{b.\sigma_y}{\sigma_x} \qquad \ldots(16)$$

Now in any random sample of n items, the average of the y.s, m_y will be $\Sigma y/n$, the average of the x.s, m_x , will be

$\Sigma x/n$. From a series of such samples, then, we would secure σ_{m_x} and σ_{m_y}. The measure of the regression between the averages would be unchanged, b, for since each $x' = b \cdot y$, each $\Sigma x'/n = b \cdot \Sigma y/n$.

Thus, just as $R_{xy} = b\frac{\sigma_y}{\sigma_x}$ so

$$r_{m_x m_y} = b\frac{\sigma_{m_y}}{\sigma_{m_x}} \qquad \ldots(17)$$

But $\sigma_{m_y} = \frac{\sigma_y}{\sqrt{n}}$ in probability))

and $\sigma_{m_x} = \frac{\sigma_x}{\sqrt{n}}$) ...(18)

Substituting in (17),

$$r_{m_x m_y} = b\frac{\sigma_y}{\sigma_x} = R_{xy} = r_{xy} \qquad \ldots(19)$$

showing the identity of correlation in the sample and correlation of samples' averages.

The reasoning in the foregoing showing the equality of the correlation in the original items in a series and in a series of averages of the items, is based on the assumption that the relationship between the variates is strictly linear. In the event that a curvilinear relation exists the correlation of the averages will nearly always be markedly higher than the correlation of the original items, as may be reasoned thus:

(a) There is practically always a grouping of the observations in a dot chart along the central portion of the regression line, with the number of cases thinning towards the extremes. If curvilinear, a curve may be drawn through them.

(b) A straight line will usually fit the central portion of the curve considerably better than it will fit the entire length since the extremes largely define the curvilinear relation.

(c) In the process of averaging a series of successive samples taken from the distribution along the curve, the central portion of the original curve is given a great deal more weight owing to the concentration of items there, than are the extremes, and thus the line of averages so secured and plotted tends to resemble the central portion of the original curve, to which a straight line may be fitted.

(d) Therefore, when dealing with data whose relation in the original is curvilinear, the correlation of the averages--correlation being in terms of linear measurement--tends to be greater than the correlation in the original data.

To ob_viate this difficulty in estimating correlation of averages in a series of successive random samples from the correlation of the unit values in the given sample, we could proceed as follows:

Supposing a curvilinear relationship--as frequently is the case--we may eliminate the effect of this in reducing correlation in the original by arbitrarily curving the regression line on the dot chart so as to pass through the greatest number of dots; measuring the standard deviations (error) of the differences around this line (in the direction of the co-ordinate representing the dependent), s, and comparing with the original standard deviation of the variable (dependent), x, gives us the correlation index $\rho_{x.y}$, (rho), by the formula of form familiar in ordinary correlation methods.

$$\rho_{x.y} = \sqrt{1 - \frac{s^2}{\sigma_x^2}} \quad \ldots(20)$$

This process is equivalent to shifting the dots on the chart to a position establishing a linear relationship. If the dots had been so moved the correlation of the original items, r'_{xy}, would be increased from r_{xy}, to equal $\rho_{x.y}$, (rho); and also the correlation r'_{xy} would equal the correlation of the averages, R_{xy}, since the curvilinearity would have been eliminated, ~~and~~ the reasoning in the foregoing portion of this section would be applicable. We may thus take $\rho_{x.y}$ instead of r_{xy} as a better indication of R_{xy}. This is basic and should be observed in similar generalizations of correlation of averages of successive samples from correlation of variates within the single sample.

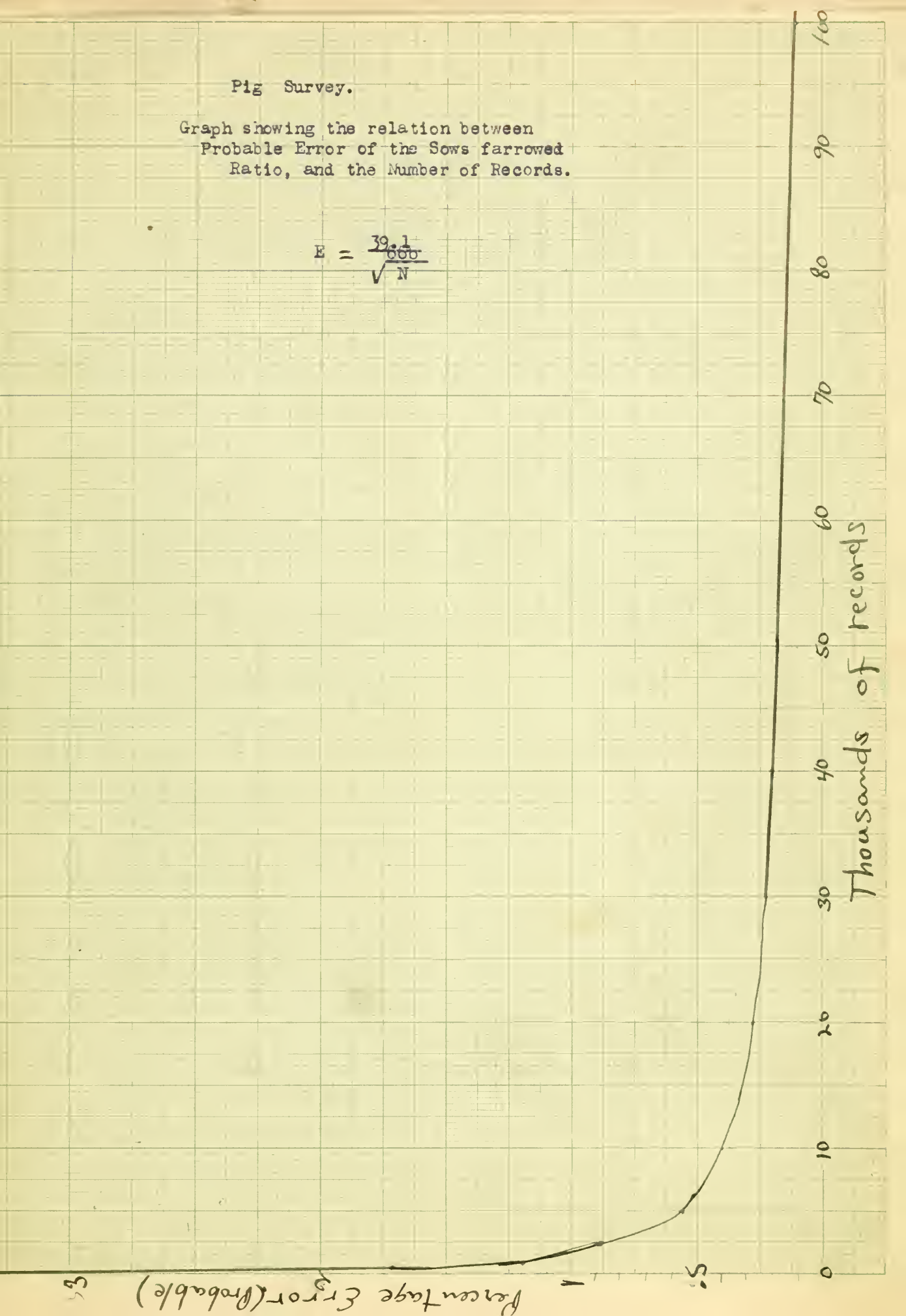
Note: There are two correlation indices just as there are two correlation ratios. But the indices have a greater tendency to coincide in magnitudes, since they are derived from "smoothed" functions avoiding the irregularities which may influence the ratios. As a rule therfore, if approximations are desired, it is only necessary to calculate one.

By refering to pages & the close agreement between $r_{m_x m_y}$, correlation of the averages and $\rho_{x.y}$, correlation ratio in the original items, in actual practice may be observed: .76 and .8 respectively.

Pig Survey.

Graph showing the relation between Probable Error of the Sows farrowed Ratio, and the Number of Records.

$$E = \frac{39.1}{\sqrt{N}}$$



COPIED

JOB NO: 11357
KIND OF WORK: Trunk (50)
NAME: Mrs. Perry

1.9	Smith, Bradford B.
Ec752Si	Size of sample study with particular reference to the pig survey... Oct.30, 1934.
JAN 18 1926	Main to be cat.
	Cat. Room
Aug.5'26	AUG 21 1926

GPO 8—2432